2010 International Conference on Pattern Recognition

Regression-based Multi-View Facial Expression Recognition

Ognjen Rudovic[1], Ioannis Patras[2], Maja Pantic[1,3]

[1]Department of Computing, Imperial College London, London SW7 2AZ, UK
{*orudovic,m.pantic*}*@imperial.ac.uk*
[2]Department of Electronic Engineering, Queen Mary University, London, UK
i.patras@eecs.qmul.ac.uk
[3]EEMCS, University of Twente, 7500 AE Enschede, The Netherlands

Abstract

We present a regression-based scheme for multi-view facial expression recognition based on 2−D geometric features. We address the problem by mapping facial points (e.g. mouth corners) from non-frontal to frontal view where further recognition of the expressions can be performed using a state-of-the-art facial expression recognition method. To learn the mapping functions we investigate four regression models: Linear Regression (LR), Support Vector Regression (SVR), Relevance Vector Regression (RVR) and Gaussian Process Regression (GPR). Our extensive experiments on the CMU Multi-PIE facial expression database show that the proposed scheme outperforms view-specific classifiers by utilizing considerably less training data.

1. Introduction

Facial expression recognition has attracted significant attention because of its usefulness in many applications such as human-computer interaction, face animation and analysis of social interaction [10]. While most existing methods focus on near frontal view images, multi-view facial expression recognition remains a significant research challenge.

Methods addressing this problem can be divided into 3D- and 2D-based methods (see [4]). The downside of the 3D-based methods is that they are computationally expensive and may fail to converge. The 2D-based methods train/apply view-specific classifiers where the number of classes increases proportionally with the number of different views and facial expressions [3]. This however increases the demand for the training data in terms of facial expressions in different views as the view-specific classifiers should be trained with a similar amount of data in order to avoid a bias [1]. However, there is a disproportion in the availability of the frontal-view and multi-view facial expression data [2, 3]. This makes it possible to build a frontal classifier by utilizing a large amount of the training data and, consequently, obtain a lower generalization error compared to non-frontal classifiers.

In this paper we propose a regression-based scheme for extending near-frontal to multi-view facial expression recognition systems. To this aim, we learn mapping functions to predict the locations of facial landmarks in the frontal view given the locations of the landmarks in the non-frontal view test images. Then, the facial expression recognition is performed by classifying the predicted landmarks' locations using a frontal classifier (see Fig. 1). This scheme allows using an existing near-frontal facial expression recognition system since the learning of the mapping functions can be carried out as an independent task.

For learning the mapping functions, we investigate the sate-of-the-art regression models: LR [1], SVR [1], RVR [8] and GPR [8]. Overall, the contributions of this paper can be summarized as follows:

1. We show that only a small amount of images with facial expressions in different views is needed to learn the target mapping functions.
2. Our experiments show that with the proposed scheme we achieve better performance than when using view-specific classifiers.
3. We compare state-of-the-art regression models and discuss their pros and cons for learning the target mapping functions.

The rest of the paper is organized as follows. In Sec. 2 we give an overview of the proposed scheme. In

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.1001

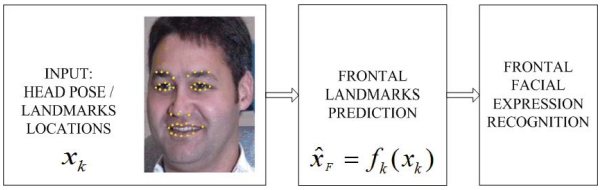

Figure 1. Outline of the proposed scheme.

Sec. 3 we discuss the regression models used for the experiments in this paper. In Sec. 4 we present and discuss the experimental results. Sec. 5 concludes the paper.

2. System Overview

Given an input image of a person exhibiting a facial expression in an arbitrary head pose, the first step is to estimate the head pose and assign it to one of P previously defined discrete poses. The second step is localization of the 2-D locations of 39 facial landmark points (see Fig. 1). In this work we assume that these steps are performed with some state of the art methods (e.g. [6], [5]) so that for an input image we know the pose k and the location x_k of the facial landmarks. Our goal is to learn functions $f_k : R^{78} \rightarrow R^{78}$ that map x_k to x_F, where x_F is the location of the facial points in the frontal view and $k = 1..P-1$. We use N training images for each of P poses to learn the functions f_k by forming pairs (x_k^i, x_F^i), where $i = 1..N$, and applying the regression models explained in Sec. 3. Once these mapping functions are learned, we use them subsequently to predict the locations of the landmark points in the frontal view $\hat{x}_F$ given the location of points x_k in the non-frontal test image. Such obtained locations are later fed into a frontal classifier that outputs the facial expression label.

3. Regression Models

This section provides a brief comparison of the regression models used to learn the above defined mapping functions f_k. LR is a parametric model which parameters are estimated using sum-of-least squares criterion [1]. These parameters can be computed only if the number of training examples $N > D + 1$, where $D = 78$ is the dimension of the input space (in our case, this is also the dimension of the output space). This imposes a serious limitations when working with high-dimensional data. This model does not suffer from local minima, and making predictions requires simple matrix multiplication. SVR [1] is a sparse kernel technique that selects M examples from a training set known as support vectors and use them to make predictions. The hyper-parameters (hp) of this model: error/margin trade-off (C), insensitivity (ϵ) and kernel parameters (θ), are usually estimated using the cross-validation technique that can be cumbersome and slow. In contrast to the above described models, GPR [8] have a probabilistic formulation and provide a principled way of model fitting by maximizing log marginal likelihood with respect to θ [1]. However, it can be easily trapped in local minima since this function is non-convex. In practice, these caveats do not prevent GPR from giving useful results. Moreover, by placing a Gaussian process prior over f_k, when making predictions, we marginalize over all possible choices of f_k, and make the model less prone to overfitting/underfitting. RVR [9] is a sparse representation of GPR with additional hp: weight (α) and noise (β) precision. These hp determine M prototypical examples called relevance vectors that are used during the prediction. Consequently, the computational cost of making predictions with GPR is typically much higher than with RVR. Compared to SVR, RVR maintains comparable generalization error while utilizing considerably lower M [1].

For SVR, RVR and GPR, we applied the kernel function: $k(x_i, x_j) = \theta_1 \exp(-\frac{\theta_2}{2}(x_i - x_j)^T(x_i - x_j)) + \theta_3 x_i x_j^T$ [1, 8]. Here, (θ_1, θ_2) are the parameters of Radial Basis Function (RBF) kernel and θ_3 corresponds to a parametric model that is a linear function of the input variables. These hp are stored in $\theta = \{\theta_i\}_{i=1}^3$. We chose this kernel function as it performed best in a pilot study compared with single RBF, MLP and Linear kernel, as well as their compound versions.

Table 1 summarizes the above discussed regression models in terms of storage and time complexity, the number of hp and the convexity of the optimization function. *Initialization* stands for the time needed to carry out preliminary matrix computations before the testing is commenced. *Prediction* refers to the time needed to make the prediction given a test datum. For all the models except of LR, we modeled the outputs independently. This resulted in a high computational load in the case of SVR, RVR and GPR. In the case of LR, the computational complexity does not depend on the number of training samples but on the number of input dimension D only. *Storage* stands for storage complexity of the discussed models.

4. Experiments

The evaluation study was conducted on the CMU Multi-PIE database [2]. Overall, 800 images across 114 subjects (74 males), with face area of approximately

Table 1. Comparison of the four regression models used in the experiments.

	Storage	Initialization	Prediction	hp	Convex
LR	$\mathcal{O}(D^2)$	$\mathcal{O}(D^3)$	$\mathcal{O}(D)$	-	yes
SVR	$\mathcal{O}(DM^2)$	$\mathcal{O}(DM^3)$	$\mathcal{O}(DM^2)$	$\{\theta, C, \epsilon\}$	yes
RVR	$\mathcal{O}(DM^2)$	$\mathcal{O}(DM^3)$	$\mathcal{O}(DM^2)$	$\{\theta, \alpha, \beta\}$	no
GPR	$\mathcal{O}(DN^2)$	$\mathcal{O}(DN^3)$	$\mathcal{O}(DN^2)$	$\{\theta\}$	no

200x250 pixels, were manually annotated with 39 landmarks (see Fig. 1). There were 200 images from each of the 4 views - frontal view, 15° left, 30° left, and 45° left - depicting 4 facial expressions (50 images per expression: Neutral (NE), Disgust (DI), Happiness (HA) and Surprise (SU)). The training data were registered as proposed in [7]. Three different mapping functions were learned for each regression method, each one for predicting the location of the facial landmarks in the frontal pose from poses 15° Left, 30° Left, and 45° Left. We used two performance measures: Error Rate (Err) in percent to report the error in the recognition of facial expressions, and Standardized Mean Squared Error (SMSE) to report the error in the prediction of the location of the facial landmarks. In Tables 2 and 3 we report the mean and standard deviation of these values obtained in the experiments explained below. For the classification of facial expressions we used support vector machine (SVM) classifier with linear kernel as it is one of the commonly used classifiers for the target problem [10].

Table 2 shows the performance of the evaluated regression models and the average performance of view-specific classifiers - the frontal-view classifier (FV-C) and the non-frontal-view classifiers (NFV-C). To evaluate the performance of various regression-based schemes and view-specific classifiers when using different amount of training data, we used N = 30, 60, 90, 120, 180. In the case of regression-based schemes, the facial expression recognition is performed by classifying the landmarks' locations predicted by the given regression model using a frontal classifier (FC), which in all cases was trained using N=180 frontal-view expressive face images. In all experiments, we applied 10-fold person-independent cross-validation. Note that RVR was able to 'discard' the noisy examples from the training set achieving better generalization ability than was the case with the other models. For example, RVR+FC has Err of 2.5% for N = 60, while GPR+FC and SVR+FC have Err of 3.33% and 5.5%, respectively. For the same case (N = 60), LR model could not be learned because N¡D. However, a general observation from Table 2 is that the regression-based schemes on average outperforms NFV-C. This is especially so for RVR and GPR, which are the linear smoothers [8], i.e. their output is a linear combination of the target (frontal) data which makes it easier for the FC to classify the predictions. GPR outperformed other regression models in terms of SMSE, meaning that it achieved more accurate prediction of the points (altough the classification of these by FC did not result in the lowest Err). We ranked the evaluated schemes according to their average Err: RVR+FC>GPR+FC>LR+FC>NFV-C>SVR+FC. It is interesting to notice that the Err for the classification of non-frontal facial expressions using either the regression-based schemes or NFV-C was lower than the Err of FV-C trained/tested using frontal data. This confirms the observation in [3] that frontal view may not be optimal for facial expression recogniton.

Table 2. Average performance results when using N training data per pose exhibiting all 4 facial expressions.

	N	30	60	90	120	180
LR+FC	Err	-	-	15.5±7.44	5.67±4.69	2.33±2.86
	SMSE	-	-	2.91±1.34	2.23±0.67	0.94±0.58
SVR+FC	Err	23.2±24.7	5.50±4.80	4.50±4.02	2.83±3.39	2.83±2.64
	SMSE	0.60±0.22	0.44±0.06	0.41±0.06	0.39±0.06	0.37±0.06
	M	23.2	46.1	64.8	81.0	119
RVR+FC	Err	17.5±23.6	2.50±3.15	2.33±3.14	2.17±3.13	2.00±2.67
	SMSE	0.62±0.21	0.45±0.08	0.42±0.08	0.39±0.06	0.37±0.06
	M	21.2	38.1	44.2	11.1	12.9
GPR+FC	Err	18.7±23.3	3.33±3.33	3.17±3.59	2.50±3.41	2.33±2.61
	SMSE	0.59±0.21	0.43±0.07	0.40±0.06	0.38±0.06	0.36±0.06
NFV-C	Err	10.8±4.48	6.17±1.15	3.83±1.04	3.50±1.50	2.67±1.89
FV-C	Err	12.0±8.23	5.55±4.97	5.55±4.38	6.00±3.94	4.50±2.84

In Table 2 we also report the average number of support/relevance vectors (M) used by SVR and RVR. Due to high dimensionality of the input space (D=78), M decreases rapidly when N¿90, thus RVR significantly outperforms SVR and GPR in terms of storage complexity. This feature of RVR is important when dealing with large datasets.

To further evaluate the robustness of the regression-based schemes in the case of missing data (i.e when no examples of a certain facial expression category were used to train the given regression model), we applied leave-one-expression-out strategy. For instance, we used the examples of NE, DI and SU to learn the target mapping functions and we used examples of HA to test those. For all evaluated regression-based schemes, the FC was trained using data of all four facial expressions. Once again we applied ten-fold person-independent cross-validation. As can be seen from Table 3 LR generalizes better on unseen data (Err$<30\%$) compared to other models. This indicates a general drawback of the kernel-based methods refering to their poor extrapolation ability [1]. As can be seen from Table 3 and 2 there is a large inconsistency between estimated SMSE and Err. Even though SMSE can be low,

at the same time Err can be very high. This implies that SMSE is not reliable error measure for the target task as it does not take into account whether the facial shape defined by the predicted points is well preserved or not.

Table 3. Results of leave-one-expression out experiment.

		Neutral	Happiness	Disgust	Surprise
LR+FC	Err	29.3±22.7	20.6±13.3	28.7±23.3	16.0±15.2
	SMSE	1.68±0.66	2.23±2.80	1.82±0.74	3.28±1.68
SVR+FC	Err	64.0±25.9	52.7±28.5	56.0±19.9	58.7±28.7
	SMSE	0.87±0.24	1.04±0.67	1.09±0.47	1.47±0.70
RVR+FC	Err	33.3±25.4	20.7±23.2	36.0±25.4	24.0±17.7
	SMSE	0.84±0.25	1.00±0.76	1.06±0.40	1.45±0.65
GPR+FC	Err	53.3±28.9	32.0±28.1	44.7±24.5	48.7±29.1
	SMSE	0.85±0.26	1.00±0.82	1.04±0.42	1.37±0.69

In real-world applications, automatically localized landmarks will inevitably be corrupted by noise. Hence, to test the robustness of the regression-based schemes in case of noisy data, we tested the models trained on noise-free data on data with added noise sampled from UNIF∼[-a,+a]. The noise level was set to $a = \sigma, .., 5\sigma$, where the standard deviation of the inputs σ was scaled to one. As can be seen from Fig. 2, SVR, RVR and GPR performed similarly (e.g., for a=0.5, Err=5%), while LR showed high sensitivity to noise (e.g. for $a = 0.5$, Err=24%).

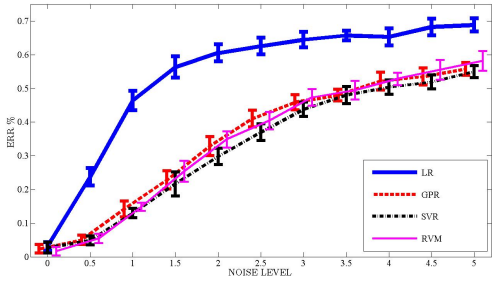

Figure 2. Err of the 4 different regression models when tested on noisy data.

5. Conclusion

In this paper, we proposed a regression-based scheme for multi-view facial expression recognition. We compared four state-of-the-art regression models within this scheme: LR, SVR, RVR and GPR. LR proved to be very sensitive to noise which makes it impractical for real-world applications. RVR and GPR performed well even when a small amount of training data were available, while SVR showed such a performance only when using a relatively large number of training data. Furthermore, parameter estimation in SVR makes it far less suitable for the target task than is the case with RVR and GPR. RVR and GPR, when applied in our scheme, outperformed view-specific classifiers. Hence, in contrast to view-specific approaches, RVR/GPR-based scheme enables multi-view expression recognition system to be designed by using considerably less training data. However, RVR performed better than GPR on the given dataset. In this paper we also showed that SMSE is not a reliable performance measure for the target task as it does not reflect accurately to what extent the information about the facial configuration is lost or preserved.

Acknowledgments

The work of Ognjen Rudovic is funded in part by the European Communitys 7th Framework Programme [FP7/2007-2013] under grant agreement no. 211486 (SEMAINE). The work of Maja Pantic is funded in part by the European Research Council under the ERC Starting Grant agreement no. ERC-2007-StG-203143 (MAHNOB). The work of Ioannis Patras is supported by EPSRC project EP/G033935/1.

References

[1] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, October 2007.

[2] R. Gross, I. Matthews, J. F. Cohn, T. Kanade, and S. Baker. Multi-pie. In *FG*, pages 1–8, 2008.

[3] Y. Hu, Z. Zeng, L. Yin, X. Wei, J. Tu, and T. Huang. A study of non-frontal-view facial expressions recognition. In *ICPR*, pages 1–4, 2008.

[4] S. Kumano, K. Otsuka, J. Yamato, E. Maeda, and Y. Sato. Pose-invariant facial expression recognition using variable-intensity templates. *IJCV*, 83(2):178–194, 2009.

[5] M.F.Valstar, B.Martinez, X.Binefa, and M.Pantic. Facial point detection using boosted regression and graph models. In *CVPR*, 2010.

[6] E. Murphy-Chutorian and M. M. Trivedi. Head pose estimation in computer vision: A survey. *TPAMI*, 31(4):607–626, 2009.

[7] M. Pantic and I. Patras. Dynamics of facial expression. *TSMC-B*, 36:433–449, 2006.

[8] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2005.

[9] M. E. Tipping. The relevance vector machine. In *NIPS*, pages 652–658, 1999.

[10] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods. *TPAMI*, 31(1):39–58, 2009.